Machine-Aided Human Translation and the Paradigm Shift

Pierre Isabelle
Canadian Workplace Automation Research Center
1575 Chomedey Blvd., Laval, Quebec, Canada H7V 2X2
e-mail: isabelle@citi.doc.ca

1. The speed at which MT researchers are abandoning or supplementing their traditional rule-based paradigms in favor of corpus-based approaches provides a striking proof of their general lack of enthusiasm for the results of the last three decades of MT research. No doubt, this feeling is fueled in part by the fact these results have not allowed the MT business to penetrate any more than a marginal segment of the global translation market.

2. Martin Kay [3] was already contemplating this failure when, in 1980, he forcefully urged the MT community to shift its near-term goal away from classical MT towards machine-aided human translation (MAHT). While many MT researchers probably felt that Kay's proposed program was a more sensible way to go, most of them declined to sign up for the journey and preferred to stick it out for another decade on their familiar path towards an improbable place.

3. This reluctance to give any serious consideration to MAHT clearly had something to do with the then-prevalent rule-based paradigms of MT. It is in fact very hard to single out, within three decades of rule-based MT research, a single result that has any obvious and immediate potential from the point of view of MAHT. There was just no straightforward way of applying the core techniques of rule-based MT systems to the development of translation support tools.

4. But just the opposite appears to be true of corpus-based approaches. For example, one of the very first results obtained within this new paradigm is the development of algorithms capable of 'aligning' the sentences of bilingual texts. This simple result turns out to be of fundamental importance from the point of view of MAHT. It constitutes in itself a suitable foundation for many kinds of new translation support tools. More on this below.

5. Why should there be such a difference between the two paradigms? The explanation, I think, is as follows. Rule-based MT tends to focus exclusively on the *translation production* problem. In the rare cases where it is possible to define good and complete translation models, this approach yields effective MT systems. But in all other cases, those where MAHT is called for, it turns out to be very difficult to make any use of production-oriented models. For example, it is hard to see how the particular target text intended by some translator could be *partially* generated in advance. Corpus-based methods, on the other hand, start from translations that have already been produced by humans and seek to discover their structure, completely or partially. This analysis-oriented perspective lends itself naturally to the development of translator's aids because in MAHT the machine is not expected to produce the translations, but rather to *understand* enough about them to become helpful.

6. *Translation analysis* is the process of making explicit some or all of the translation correspondences that link the segments of a source text with those of its translation. Sentence alignment is a kind of translation analysis that aims at showing correspondences down to the level of sentences but no further. No more than an extremely weak translation model appears to be needed for that

particular task ([1], [4]). Research is currently underway to develop analyzers capable of accounting for finer translation correspondences (e.g. between phrases, words and even morphemes). But existing techniques already open the way to a whole new family of translation support tools, including: 1) translation memory applications; 2) translation checkers; and 3) translation dictation systems.

7. Translation analysis can be used to structure pre-existing translations in such a way that they become reusable in the production of new translations. For example, the CWARC's TransSearch system [2] is a *bilingual concordancing* tool that allows translators to search special-purpose databases for ready-made solutions to specific translation problems. These databases result from the application of sentence alignment techniques to pre-existing translations.

8. Translation analysis can also be applied to draft translations with the goal of detecting certain kinds of translation errors. To the extent that we can automatically reconstruct the translation correspondences between the source and a draft target text, we can verify that these correspondences obey certain constraints. For example, it should be possible to verify that the translation is complete, in the sense that all large chunks (pages, paragraphs, sentences) of the source text have been translated. It should also be possible to verify that the translation is free from interference errors such as those caused by *deceptive cognates* (e.g. English *actual* and French *actuel* are not acceptable as mutual translations). The CWARC is currently developing a prototype translation checker called TransCheck [2].

9. Finally, translation analysis can be used to improve the performance of speech recognition for dictated translations. Suppose that a translator is about to dictate his translation of an English sentence that contains the word *government*. Given the situation, the likelihood that his utterance will contain the acoustic image of the French word *gouvernement* becomes much higher than for a chance utterance. This kind of probabilistic translation knowledge can be built into a speech recognition system used for the automatic speech-to-text transcription of dictated translations. This amounts to applying translation analysis to a situation where the target text is spoken rather than written. The TransTalk project [2], a joint effort of the CWARC and the Centre de recherche informatique de Montréal (CRIM), is currently exploring this approach.

10. While there is no certainty that the current paradigm shift towards corpus-based approaches will result in better MT systems, we conclude that it will be highly productive from the point of view of MAHT. Emerging translation analysis techniques promise no less than a new generation of tools for translators. MAHT is about to become a booming business.

REFERENCES

[1] Brown P., Lai J., Mercer R., *Aligning Sentences in Parallel Corpora*, **Proceedings of the 29th Meeting of the ACL**, 1991.

[2] Isabelle P., Dymetman M., Foster G., Jutras J.-M., Macklovitch E., Perrault F., Ren X., Simard M., *Translation Analysis and Translation Automation*, **Proceedings of TMI-93**, Kyoto, 1993.

[3] Kay M., **The Proper Place of Men and Machines in Translation**, CSL-80-11, Xerox PARC, 1980.

[4] Simard M., Foster G., Isabelle P. *Using Cognates to Align Sentences in Parallel Corpora*, **Proceedings of TMI-92**, Canadian Workplace Automation Research Center, Montreal, 1992.

Machine-Aided Human Translation and the Paradigm Shift

Pierre Isabelle
Canadian Workplace Automation Research Center
1575 Chomedey Blvd., Laval, Quebec, Canada H7V 2X2
e-mail: isabelle@citi.doc.ca

1. The speed at which MT researchers are abandoning or supplementing their traditional rule-based paradigms in favor of corpus-based approaches provides a striking proof of their general lack of enthusiasm for the results of the last three decades of MT research. No doubt, this feeling is fueled in part by the fact these results have not allowed the MT business to penetrate any more than a marginal segment of the global translation market.

2. Martin Kay [3] was already contemplating this failure when, in 1980, he forcefully urged the MT community to shift its near-term goal away from classical MT towards machine-aided human translation (MAHT). While many MT researchers probably felt that Kay's proposed program was a more sensible way to go, most of them declined to sign up for the journey and preferred to stick it out for another decade on their familiar path towards an improbable place.

3. This reluctance to give any serious consideration to MAHT clearly had something to do with the then-prevalent rule-based paradigms of MT. It is in fact very hard to single out, within three decades of rule-based MT research, a single result that has any obvious and immediate potential from the point of view of MAHT. There was just no straightforward way of applying the core techniques of rule-based MT systems to the development of translation support tools.

4. But just the opposite appears to be true of corpus-based approaches. For example, one of the very first results obtained within this new paradigm is the development of algorithms capable of 'aligning' the sentences of bilingual texts. This simple result turns out to be of fundamental importance from the point of view of MAHT. It constitutes in itself a suitable foundation for many kinds of new translation support tools. More on this below.

5. Why should there be such a difference between the two paradigms? The explanation, I think, is as follows. Rule-based MT tends to focus exclusively on the *translation production* problem. In the rare cases where it is possible to define good and complete translation models, this approach yields effective MT systems. But in all other cases, those where MAHT is called for, it turns out to be very difficult to make any use of production-oriented models. For example, it is hard to see how the particular target text intended by some translator could be *partially* generated in advance. Corpus-based methods, on the other hand, start from translations that have already been produced by humans and seek to discover their structure, completely or partially. This analysis-oriented perspective lends itself naturally to the development of translator's aids because in MAHT the machine is not expected to produce the translations, but rather to *understand* enough about them to become helpful.

6. *Translation analysis* is the process of making explicit some or all of the translation correspondences that link the segments of a source text with those of its translation. Sentence alignment is a kind of translation analysis that aims at showing correspondences down to the level of sentences but no further. No more than an extremely weak translation model appears to be needed for that

particular task ([1], [4]). Research is currently underway to develop analyzers capable of accounting for finer translation correspondences (e.g. between phrases, words and even morphemes). But existing techniques already open the way to a whole new family of translation support tools, including: 1) translation memory applications; 2) translation checkers; and 3) translation dictation systems.

7. Translation analysis can be used to structure pre-existing translations in such a way that they become reusable in the production of new translations. For example, the CWARC's TransSearch system [2] is a *bilingual concordancing* tool that allows translators to search special-purpose databases for ready-made solutions to specific translation problems. These databases result from the application of sentence alignment techniques to pre-existing translations.

8. Translation analysis can also be applied to draft translations with the goal of detecting certain kinds of translation errors. To the extent that we can automatically reconstruct the translation correspondences between the source and a draft target text, we can verify that these correspondences obey certain constraints. For example, it should be possible to verify that the translation is complete, in the sense that all large chunks (pages, paragraphs, sentences) of the source text have been translated. It should also be possible to verify that the translation is free from interference errors such as those caused by *deceptive cognates* (e.g. English *actual* and French *actuel* are not acceptable as mutual translations). The CWARC is currently developing a prototype translation checker called TransCheck [2].

9. Finally, translation analysis can be used to improve the performance of speech recognition for dictated translations. Suppose that a translator is about to dictate his translation of an English sentence that contains the word *government*. Given the situation, the likelihood that his utterance will contain the acoustic image of the French word *gouvernement* becomes much higher than for a chance utterance. This kind of probabilistic translation knowledge can be built into a speech recognition system used for the automatic speech-to-text transcription of dictated translations. This amounts to applying translation analysis to a situation where the target text is spoken rather than written. The TransTalk project [2], a joint effort of the CWARC and the Centre de recherche informatique de Montréal (CRIM), is currently exploring this approach.

10. While there is no certainty that the current paradigm shift towards corpus-based approaches will result in better MT systems, we conclude that it will be highly productive from the point of view of MAHT. Emerging translation analysis techniques promise no less than a new generation of tools for translators. MAHT is about to become a booming business.

REFERENCES

[1] Brown P., Lai J., Mercer R., *Aligning Sentences in Parallel Corpora*, **Proceedings of the 29th Meeting of the ACL**, 1991.

[2] Isabelle P., Dymetman M., Foster G., Jutras J.-M., Macklovitch E., Perrault F., Ren X., Simard M., *Translation Analysis and Translation Automation*, **Proceedings of TMI-93**, Kyoto, 1993.

[3] Kay M., **The Proper Place of Men and Machines in Translation**, CSL-80-11, Xerox PARC, 1980.

[4] Simard M., Foster G., Isabelle P. *Using Cognates to Align Sentences in Parallel Corpora*, **Proceedings of TMI-92**, Canadian Workplace Automation Research Center, Montreal, 1992.